European Commission

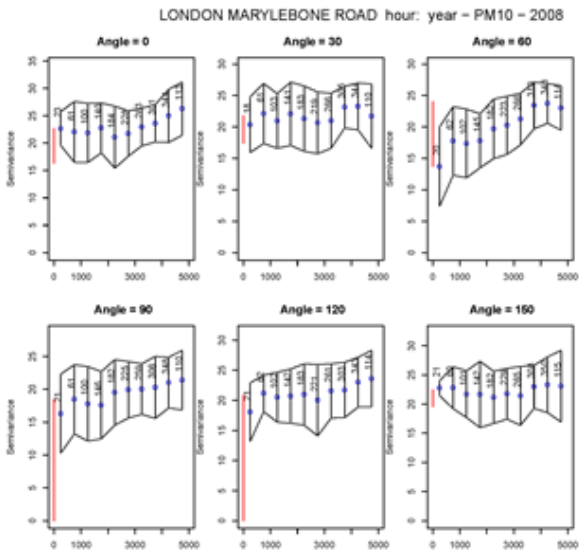
JRC TECHNICAL REPORTS

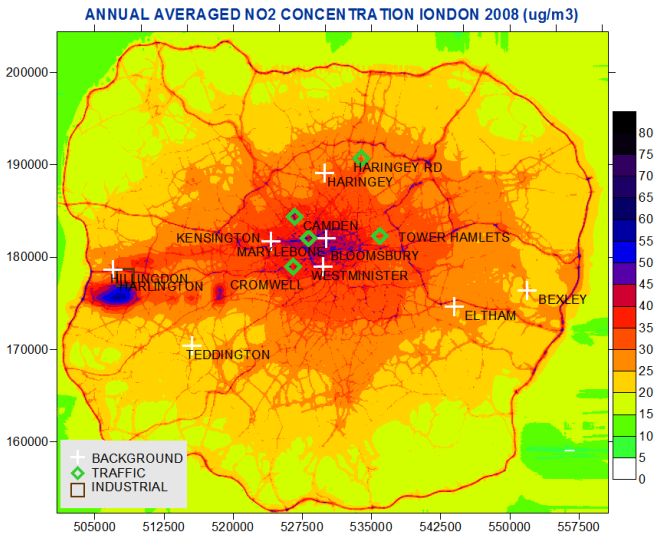
SPATIAL REPRESENTATIVITY

REPORT OF 2013 WG2/SG1 ACTIVITY

Author: Efisio Solazzo
Contributing authors:
Michel Gerboles, Oliver Kracht, Jenny Stocker*, David Carruthers* and Stefano Galmarini

2014





Report EUR 26539 EN

Joint Research Centre

European Commission
Joint Research Centre
Institute for Environment and Sustainability

Contact information
Stefano Galmarini
Address: Joint Research Centre, Via Enrico Fermi 2749, TP 441, 21027 Ispra (VA), Italy
E-mail: stefano.galmarini@jrc.ec.europa.eu
Tel.: +39 0332 795382
Fax: +39 0332 785466

(*) Cambridge Environmental Research Consultants (CERC), Cambridge (UK)

This publication is a Reference Report by the Joint Research Centre of the European Commission.

JRC87277

EUR 26539 EN

ISBN 978-92-79-35592-9 (print)
ISBN 978-92-79-35591-2 (pdf)

ISSN 1018-5593 (print)
ISSN 1831-9424 (online)

doi: 10.2788/14035

Luxembourg: Publications Office of the European Union, 2014

Executive Summary

The attached document addresses the progress of the FAIRMODE/SG1 activity since it starts in June 2013 up to the month of November 2013, for a total of about 100 working days.
During this period the analysis has focused on the development of a novel methodology – *point-centred semi-variogram* (pcsv) – for help assessing the spatial representativeness of the air quality receptors in Europe. The technique has been successfully developed and tested on a proxy of measured ambient air quality concentration data, namely results from an atmospheric modelling system.
Action taken:

1. Established collaboration with the CERC group of the UK, developer and distributor of the Atmospheric Dispersion Modelling System (ADMS), for sharing of model results of pollutant dispersion in London. These data have been used as proxy in our analysis;
2. Established collaboration with the ARIA technologies SA of France for sharing of the AIRCITY high resolution modelling results of pollutant dispersion in Paris. The ARIA group has manifested the willingness to cooperate on the SG1 activity, although data for Paris have not yet been exploited.
3. Development, testing, and coding of the pcsv technique;
4. Application of the pcsv methods to the ADMS data using the positioning of European AIRBASE air quality receptors as centre of the analysis;
5. Preparation of the attached report.

We note that about 90% of the total time has been spent on item 3 and 4 of the action list. The code we have produced is computing-intensive and has been run at the limit of the available computing resources (single PC unit). The code requires about ten days to complete a full run (three species, two years, fourteen central points).
The results obtained are promising, the technique is robust and applicable to several pollutants, though it should be considered that generalisation is difficult, as any other cases needs to be treated individually. The necessity of developing a generalised method will require more time as other cases will need to be considered and the analogies among them taken in to account.

S. Galmarini
(coordinator of the FAIRMODE/SG1)

The main aim of SG1 is to find a formulation of spatial representativeness that could be acceptable to both the monitoring and the modeling communities that considers all physical elements determining the representativeness and that is manageable in the context of monitoring and modeling data treatments.

1. INTRO AND SCOPES

The assessment of spatial representativeness of air quality monitoring stations is an outstanding issue that impinges on several areas relevant to risk assessment and population exposure as well as on the design of monitoring networks, model development, evaluation and data assimilation. There are several approaches proposed in the literature that try to define the area of representativeness of a monitoring station as "a similar area" or "spatial homogeneous field of pollution" (e.g. Bobbia et al., 2008). Such a definition cannot fit the intrinsic anisotropy of the atmospheric flow and dispersion, and is limited in time.

In the kick-off meeting of the SG1 activity held in the 28th of June 2013, it was discussed the option to develop a *point-centered semi-variogram* (*pcsv*) technique for a number of regulated pollutants measured by ground-level receptors. Variogram analysis is a well-established geostatatistical tools used to describe the spatial variance across a given region. All possible station pairs are examined by considering their mutual pair-wise distance within a given region, according to the following relationship:

$$\gamma(h) = \frac{1}{2N(h)} \sum_{h_{ij}} \left(x_i - x_j\right)^2 \qquad j = 1, \ldots, n_R, \ \forall j \neq i \quad (1)$$

where $N(h)$ is the number of station pairs separated by a distance (or lag) of h and n_R is the number of points in the region. However, as it has been pointed out by e.g. Janis and Robeson (2004) and others, this method is not suitable for assessing the representativeness of single monitoring point. The *pcsv* technique we propose here is a modification of the standard method, and involves a semi-variogram constructed by relating one point (x_0, the central measurement point) to the set of other points within a radius R, with the scope of assessing how the collection of neighboring points relates to that single measurement point:

$$\gamma_{x_0}(h) = \frac{1}{2N(h)} \sum_{(j)|h_{0j}=h} \left(x_0 - x_j\right)^2 \qquad j = 1, \ldots, n_R \quad (2)$$

where n_R is the number of points within a distance R from x_0.

In this way the *pcsv* method excludes the mutual relation of all points with each other and considers the measurement point a center of the analyses.

Our proposal is to use the area within which the level of confidence is 95 % that the true pollutant value is included in the interval $x_i \pm V$, where x_i is the measurement value at the central point and V combines the contributions of the measurement uncertainty and uncertainty arising from the spatial variability within the area. The latter contribution is calculated from spherical models fitted to *pcsv* and their directional evolution over time. To build this type of variograms, the pairs of values shall consist of

- The concentration value at the central point, and
- values of a densely known explicative variable within a given radius of the central point.

Because of the sparse nature of real monitoring networks (low space resolution), in our case the central point is represented by the location of the air quality monitoring station of the AIRBASE station, while the explicative variable consisted of modelled concentration values of NO_2, O_3, and PM_{10}, calculated using the urbanised version of the Atmospheric Dispersion Modelling System ADMS (Carruthers et al. 2001) developed by CERC. ADMS has been run with a high resolution output grid over the city of London for the years 2008 and 2011, with a model domain exceeding 50 km by 50 km. Yearly-averaged hourly concentrations of NO_2, O_3, and PM_{10} were calculated at each grid node (further details are given in section 2.1). The center-points have been selected as the grid nodes closer to the location of the AIRBASE monitoring stations in London (Fig. 1).

The methodology outlined above is novel for two reasons:

1) it has never been applied before to air quality monitoring stations, and

2) it is used here for spatial outputs of a city-scale dispersion model.

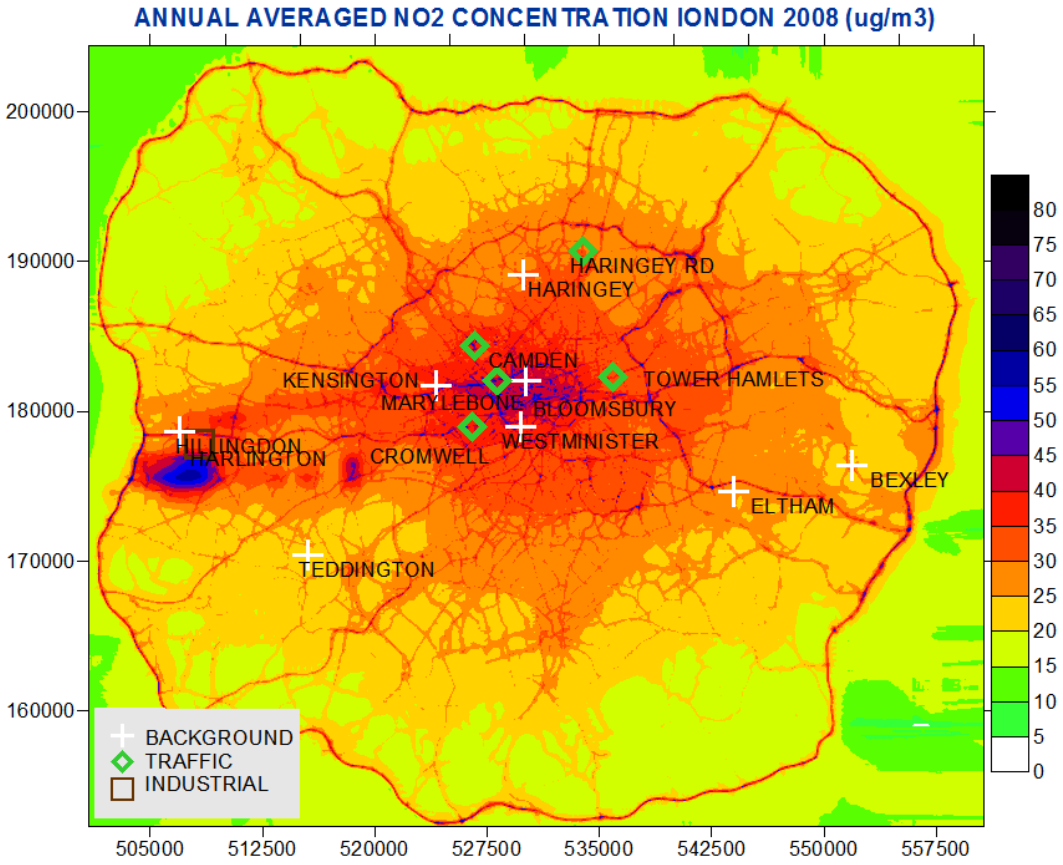


Figure 1. Gridded annual averaged NO2 concentration for the year of 2008 produced by the ADMS modeling system. Overlaid is the location of 14 monitoring stations of the AIRBASE network.

Thus, the *pcsv* technique has been applied to a proxy of measured pollutant concentration, i.e. the output of a high resolution air quality dispersion model. The gridded model outputs served as a dense explicative variable for the field of pollutant in the neighborhood of real monitoring station. Using gridded model outputs as proxy allowed focusing exclusively on the development of the *pcsv* technique rather than dealing with the issues of quality, quantity, density of the monitoring data, which are rather common in variogram analysis and require long time to address.

To date, applications of *pcsv* are scarcely documented in the scientific literature, and the technique has not been used to air quality monitoring network, or to high density proxy data, before. Furthermore, the

code underlying *pcsv* is not available in any available statistical package. **It was thus necessary to code it from scratch in the framework of the present analysis**. The material in Appendix 1 describes the steps taken so far in this direction, highlights the testing and justifies the strategy undertaken.

2. METHODS

2.1 MODELLED POLLUTANT DISPERSION OVER LONDON USING ADMS-URBAN

The Urban version of the ADMS modelling system is an advanced three-dimensional quasi-Gaussian model calculating concentrations hour by hour, nested within a straight-line Lagrangian trajectory model. The model provides predictions on a continuous surface i.e. there is no distinction between 'background' and 'roadside'. Predictions were available on a variable receptor grid with enhanced resolution close to roads. ADMS-Urban uses an 8-reaction atmospheric chemistry scheme based on Generic Reaction Set (NO, NO_2, O_3, VOC) and a simplified scheme for sulphate generation. Specific account is taken of primary NO_2 fractions. Meteorology is based on hourly surface measurements at Heathrow Airport. The boundary conditions for different species are based on hourly measurements from rural sites around London. For the current study the model produced hourly predictions of all required species including NO_2, O_3, and PM_{10}. Similarly, all scenarios were modelled, based on the London Atmospheric Emissions Inventory. Note that ADMS-Urban considers hourly profiles in emissions sources. Full description of the ADMS-Urban modelling system, documentation and evaluation studies can be found at http://www.cerc.co.uk/environmental-software/publications.html.

The grid spacing is irregular. The finest cells size if of ~50 m in the centre of the domain, reducing to between 250-300 m towards the edge of the domain. There are many additional output points adjacent to the roads, to improve output resolution in these areas. The contour map of Figure 1 shows the yearly-averaged NO_2 concentration (year 2008) produced by the ADMS-Urban model.

2.2 ASSUMPTIONS AND IMPLEMENTATION

The grid node closer to the position of each monitoring station (Table 1) is the centre of a region having radius R = 5 km. After several tests, it was decided to smooth any possible concentration gradient in the vicinity of the central point; therefore the irregular grid of the model outputs was interpolated to a regular grid with grid spacing of 80 m. The interpolation was also found necessary in light of the computing resources available: the original grid was far too dense, resulting in running time errors. The use of a regular grid for this initial analysis has allowed to not reducing the value of R. The number of points in each region was of ~200,000, the maximum found practicable.

Variogram analysis based on Eq.(2) has been performed on a number of directions, from north to south with interval of 30 degrees and tolerance of 15 degrees. Directional effects (anisotropy) is in fact anticipated, given the nature of the urban areas under examination, with long street canyons channeling the flow, and repeated urban neighborhood units. The directions analysed were: 0°± 15°, 30°± 15°, 60°±

15°, 90°± 15°, 120°± 15°, 150°± 15°, which are symmetric for 180° angle, meaning that the direction 0° includes the 0°-180° axis, the direction 30° includes the 30°-210° axis, etc. Such a symmetry is however, artificious and derives from the quadratic nature of Eq.(2). For applications to urban areas further work needs to be focus on non-symmetric directional variogram analysis.

Table 1 lists the receptors afferent the AIRBASE network within the modeled London area. The classification into (urban) background, roadside and industrial is taken from the receptor's metadata provided by AIRBASE.

Table 1 AIRBASE air quality monitoring stations in London

station name	type of station	X (m)	Y (m)
LONDON BEXLEY	Background	551859	176381
LONDON BLOOMSBURY	Background	530119	182039
LONDON CROMWELL ROAD 2	Traffic	526529	178966
LONDON ELTHAM	Background	543986	174665
LONDON HARINGEY	Background	529899	189129
LONDON HARLINGTON	Industrial	508295	177800
LONDON HILLINGDON	Background	506941	178610
LONDON MARYLEBONE ROAD	Traffic	528126	182015
LONDON N. KENSINGTON	Background	524049	181751
LONDON TEDDINGTON	Background	515545	170416
LONDON WESTMINSTER	Background	529778	178957
LONDON CAMDEN KERBSIDE	Traffic	526633	184390
LONDON HARINGEY ROADSIDE	Traffic	533909	190674
LONDON TOWER HAMLETS ROADSIDE	Traffic	535936	182271

The directional *pcsv* analysis has been performed, for each receptor, for the yearly-averaged hourly concentration of NO_2, PM_{10} and O_3 for the simulated years of 2008 and 2011. As an example, Fig. 2 shows the maps of PM_{10} concentration and squared concentration distance from the central point for the year 2008, relatively to the Marylebone roadside station. The first map pair (top right) is the yearly average, while each map after the first reflects one yearly-averaged hour. The high PM_{10} concentration values (yellow) correspond to the busy street network of London, and reflects the traffic vehicle emissions. Similar maps have been produced for all of the receptor of Table 1 taken as central point of 5 km radius region.

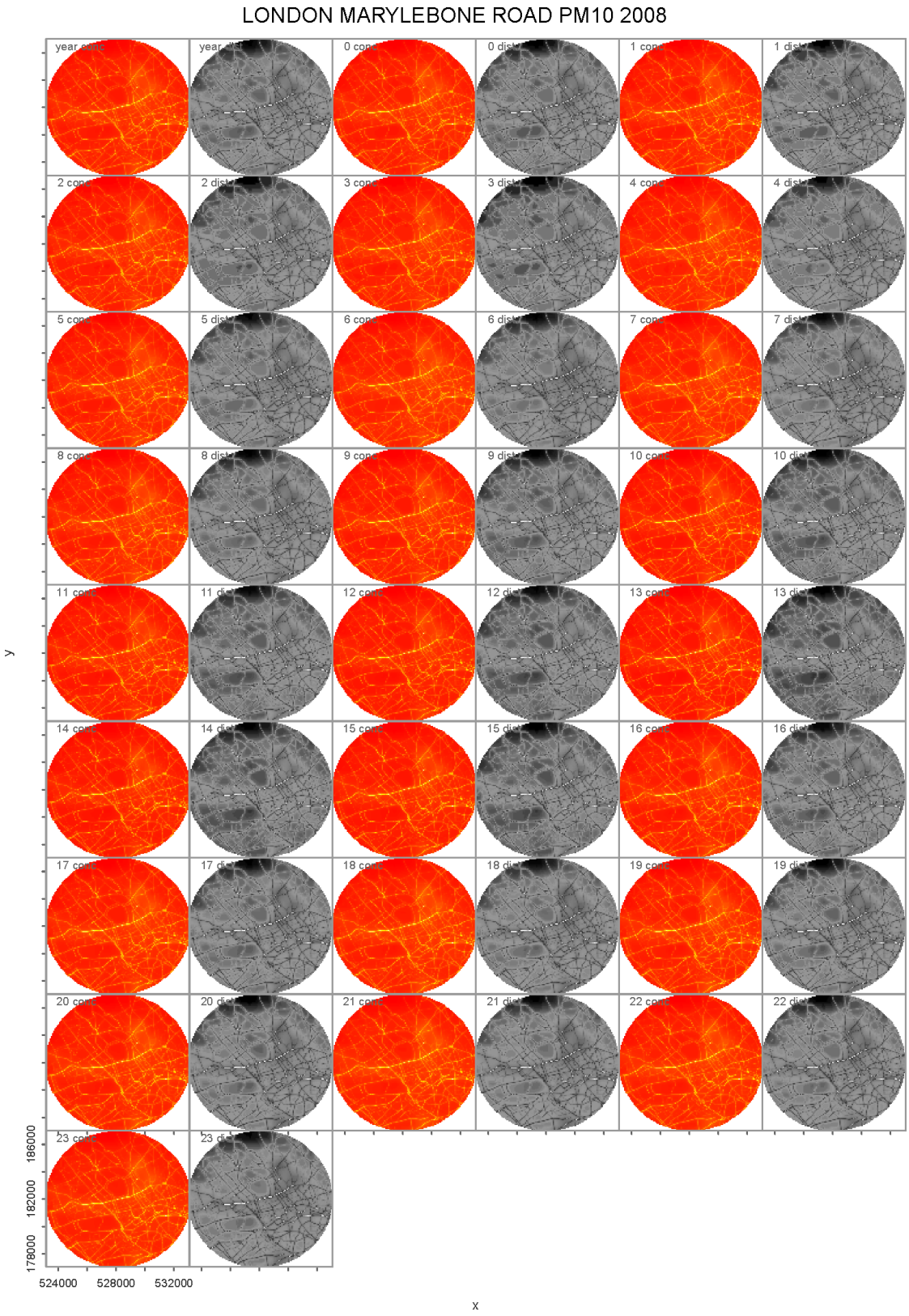


Figure 2. For each analysed time the figure shows the concentration map (min: red; max: yellow) in a 5 km radius around the location of the AIRBASE station of Marylebone RD taken as central point and the distance map (min: white; max: black) calculates as the squared difference between the concentration value at the central point and any other point.

3. VARIOGRAM MODELS

A sample variogram describes how the spatial continuity changes with distance and direction. However, for the purposes of interpolating (kriging) points where data are not available, variogram values are required also for between sample locations, and in general, at locations where an estimate is desired

(Isaaks and Srivastava, 1989). Thus a model is required enabling to compute a variogram value for any possible separation vector.

Three parameters can be derived when most functions are fit to semi-variograms: the *sill*, *range*, and *nugget* (Fig. 4).

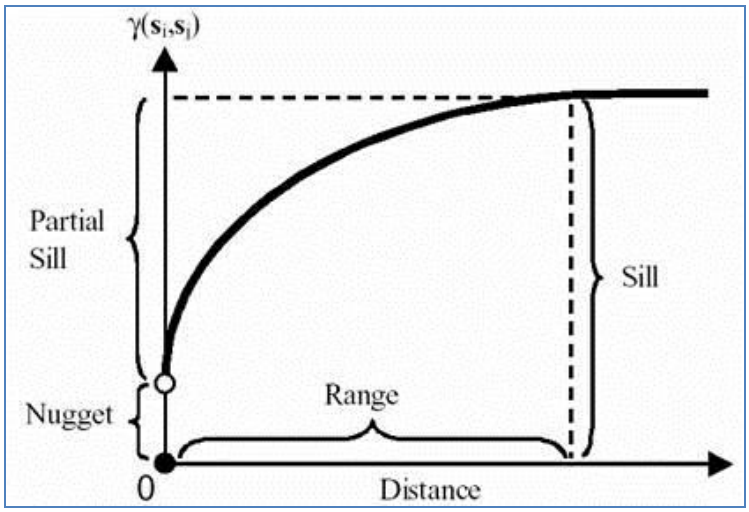


Figure 3. Parameters that can be derived from a variogram model

Although there can be considerable variation among variogram models, these parameters can be described generally (Isaaks and Srivastava, 1989). The sill is a plateau in semi-variance that occurs at a distance defined by the range. At distances greater than the range, the regionalized variable is independent (of itself).

The spherical model is one of the most widely applied variogram models, whose standardised equation is:

$$\gamma(h) = \begin{cases} 1.5\frac{h}{a} - 0.5\left(\frac{h}{a}\right)^3 & if\ h \leq a \\ 1 & otherwise \end{cases} \quad (3)$$

where *a* is the range and *h* is the lag distance. Eq. (3) has a liner behavior near the origin and flattens out for large *h*, reaching the sill for *h*=*a*. Eq. (3) yields $\gamma(0) = 0$, but the variogram value at very small separation distances may largely differ from zero, giving rise to discontinuity. This is known as *nugget effect*. The nugget is the estimated non-zero semi-variance as distance approaches zero. The nugget accounts for fine-scale variations that are unresolved by the sampling network. *The nugget of a pcsv model, therefore, can be used to estimate the degree to which the regional values approximate the target station value* (Janis and Robeson, 2004). A large nugget indicates that, over short distances, semi-variance behaves differently than the larger-scale spatial pattern would have otherwise predicted. Because nugget estimates are variance extrapolated to zero distance, a scarcity of nearby neighbors could introduce greater uncertainty into nugget estimates. The efficacy of these extrapolations is one of the potential limitations of our method.

As an example of the analysis produced, Fig. 4 reports the variogram cloud and model relatively to the yearly-averaged PM_{10} concentration hour of 7 am of the year 2008 for the Marylebone RD station. The fit parameters (sill, nugget, range) are clearly dependent on the directionality. As the main street axes is approximately in the 90°-270° direction, and secondary busy roads lie approximately in the 120°-300°

direction, it would be expected that the nugget to be small in those directions (spatial continuity in the vicinity of the station) and the range to be quickly reached. This is indeed the case as emerges from Fig. 4d and 4e. By opposite, in the instances where the concentration of the monitoring point is compared to values taken away from emission sources (obstructions, flanking buildings and/or building's facades) large discontinuity emerges in the vicinity of the monitoring point, and the nugget value is larger, as in the case of the 60°-240° direction. In this latter case the range derived by the spherical model fitting is even larger than the radius of the investigated area, indicating a drift or trend in the data: the squared concentration distance keeps increasing with distance and does not reach a plateau value within the investigated region.

That of a drift in the variogram map is a recurrent behavior noted more frequently for NO_2 and O_3 (always depending on the direction). Such a behavior may indicate that the spherical model is not adequate in instances when the concentration distances keep increasing with distance values, or that we need to extend the radius of the region to include more points and allow reaching a spatial homogeneity. Investigating the causes and solution of trend in the data is left to future investigation.

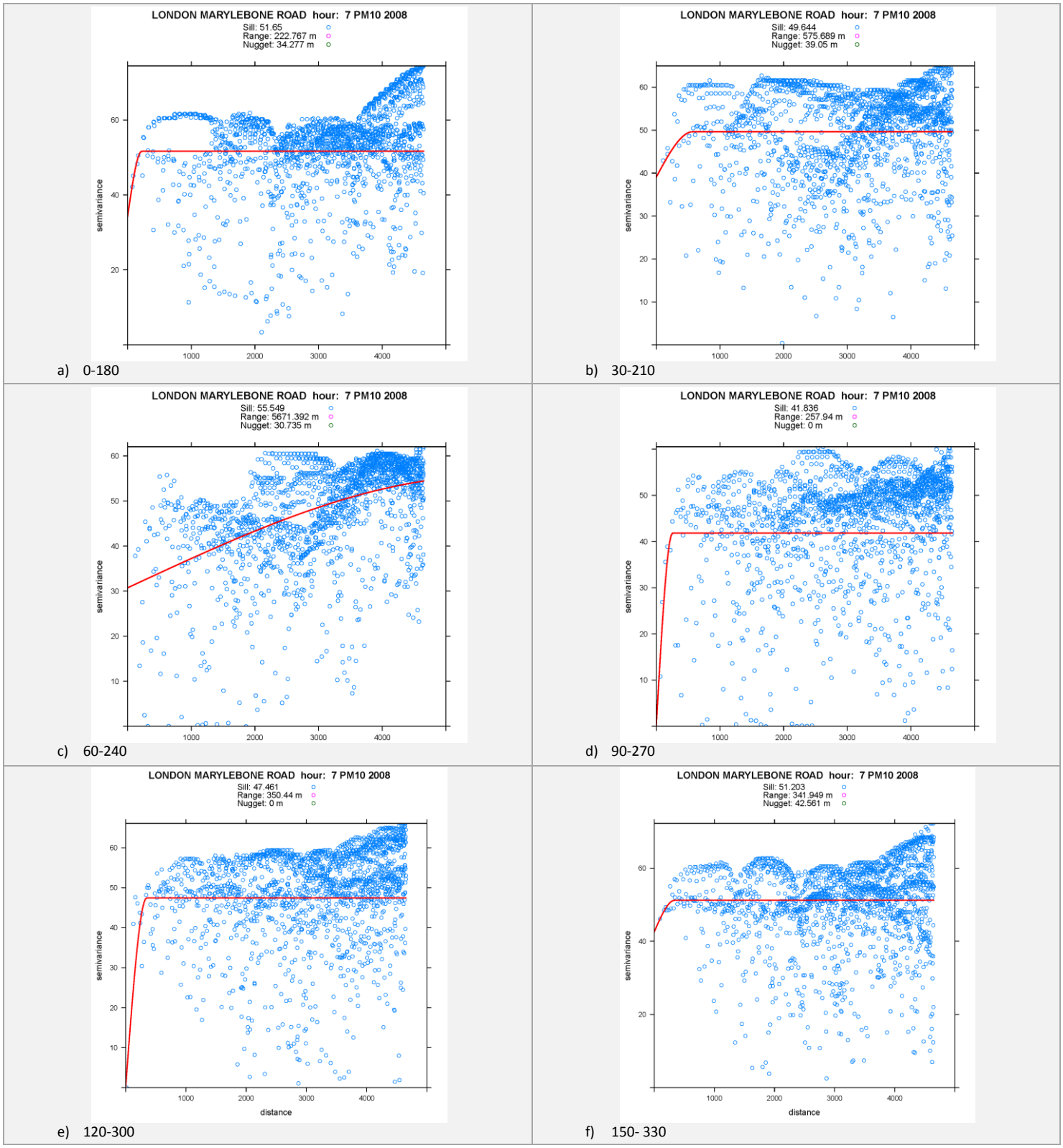


Figure 4. Directional variogram map (blue dots) and variogram model (continuous red line) as function of distance. The plots are relative to the yearly-averaged PM_{10} concentration hour of 7 am of the year 2008 for the Marylebone station. The direction is reported at the bottom, summary parameters at the top of each plot.

A further aspect needing attention is the lack of points in the vicinity of the central points in Fig. 4. Even at small distances the majority of point exhibit high variance. It remains to be established whether such a behavior is due to large concentration gradients or to a problem with the interpolation of the irregular model grid in the vicinity of the central point.

Figure 5 reports an example of the bins population with the associated standard deviation, for the yearly-averaged concentration of PM_{10} in 2008 and 2001. After several tests and literature search we decided to divide up the region into 10 bins equally spaced. The number of points in each bin is not a function of directionality, as the grid is uniform. Small variations of bins population with direction are due to the tolerance introduced when creating the variogram cloud. The variability, on the other hand, is a function of direction and distance and is more uniform for directions along the road main axes (90° and 120°) than for directions crossing urban units (street canyons, neighborhoods).

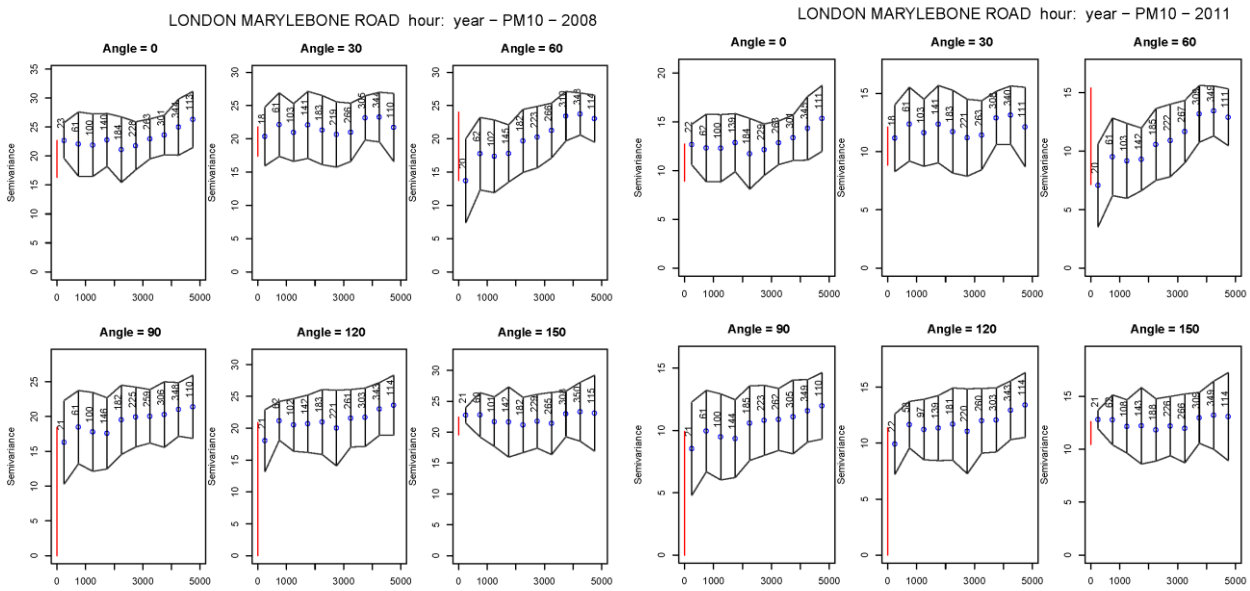


Figure 5. Bins population and variability (standard deviation) associated to each bin for different directions. Receptor: Marylebone RD; Species: PM_{10}; year: 2008 (left) and 2011 (right).

Furthermore, by comparing the variability distribution over the two years it results that while the overall shape remains the same (the street network did not change), the variability in 2011 is considerably lower than in 2008 (note the difference scales on the vertical axes for the two groups of plots), indicating a more uniformly mixed PM_{10} concentration in 2011. This is going to impact the model parameters, as discussed in the next section.

4. COLLECTIVE RESULTS

In this section the results for the model parameters (range, sill, nugget, error) are presented collectively. As illustrative example, results are shown for Marylebone RD. Similar results have been derived for all receptors of Table 1.

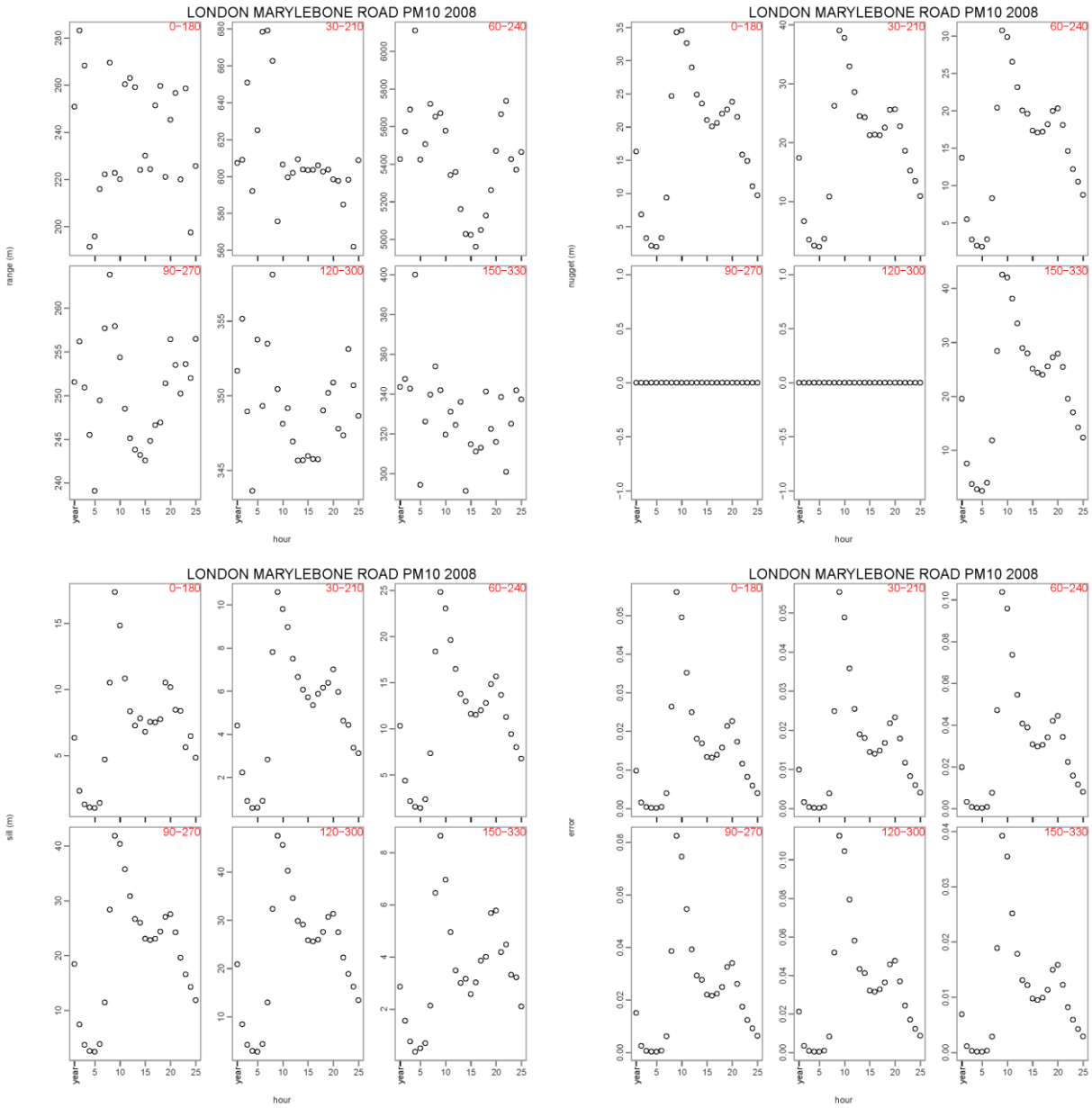


Figure 6 Collective model parameters for the receptor of Marylebone RD, year 2008, species PM_{10}. Clockwise from top right: range, nugget, error, and sill.

The plots in Fig. 6 shows the collective values of the range, the nugget, the sill, and the error of the model fit as function of the hour. The error is the mean quadratic error of the model fit across all bins. The variance of the error corresponds to the *kriging variance*, a crucial parameter for assessing the uncertainty associated with the modelling procedure.

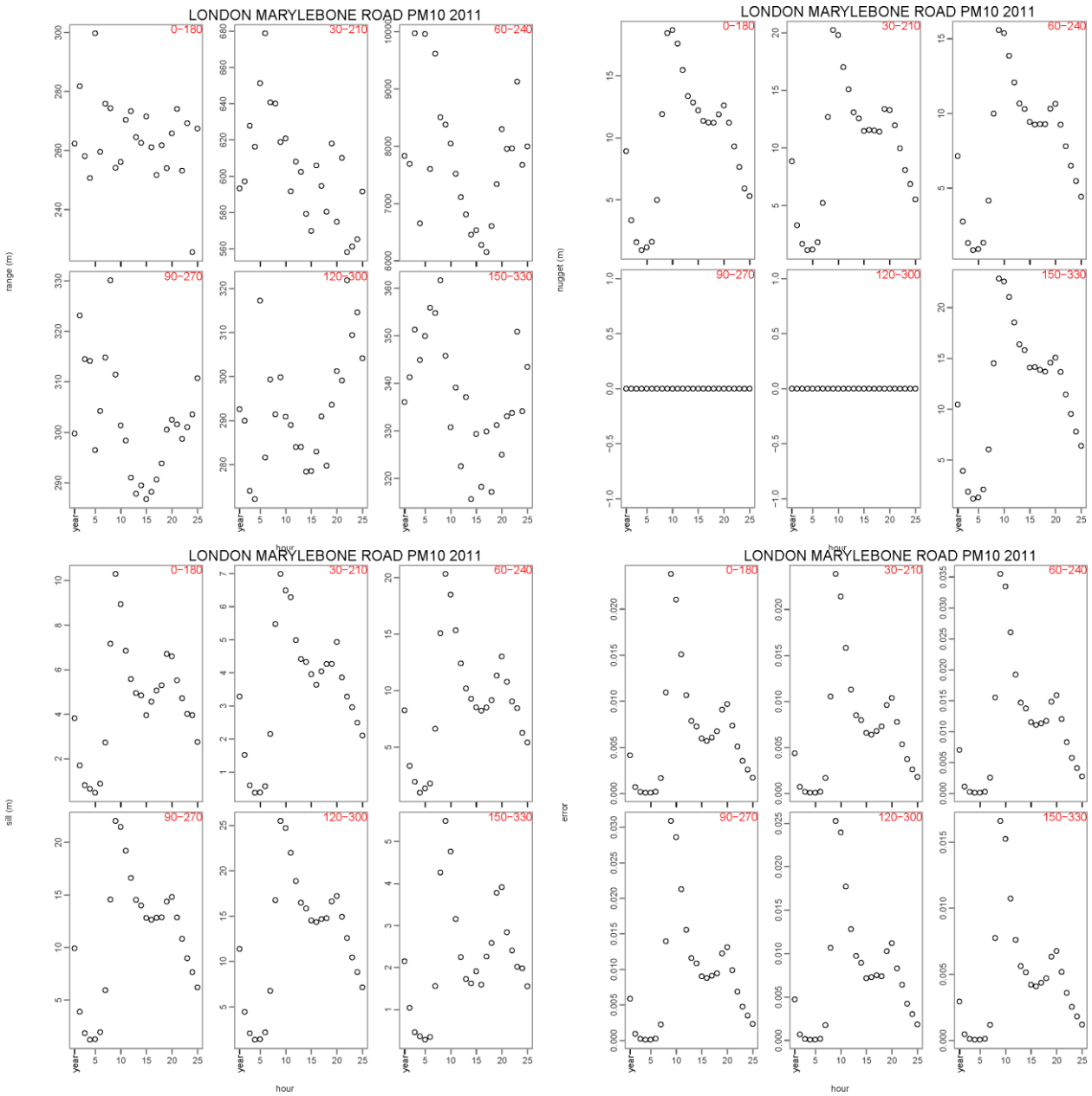


Figure 7. Collective model parameters for the receptor of Marylebone RD, year 2011, species PM_{10}. Clockwise from top right: range, nugget, error, and sill.

The values of the range for the direction 60°-240° are unphysical and confirm that the spherical model is not appropriate to model the trend of the data for that region. Overall, the range for PM_{10} (year 2008) is between 200 and 700 m, while the nugget between approximately 0 and 40 m. The small values of the error indicates that there is very little variance associated to the calculation of the fitting parameters for PM_{10}.

Thanks to the ADSM modelling data provided by CERC, we can compare the results for the same species over two different years and also between different species. In Fig. 7 and Fig. 8 we propose the same collective analysis of Fig. 6 for PM_{10} averaged over the year 2011 and for NO_2 for the same year.

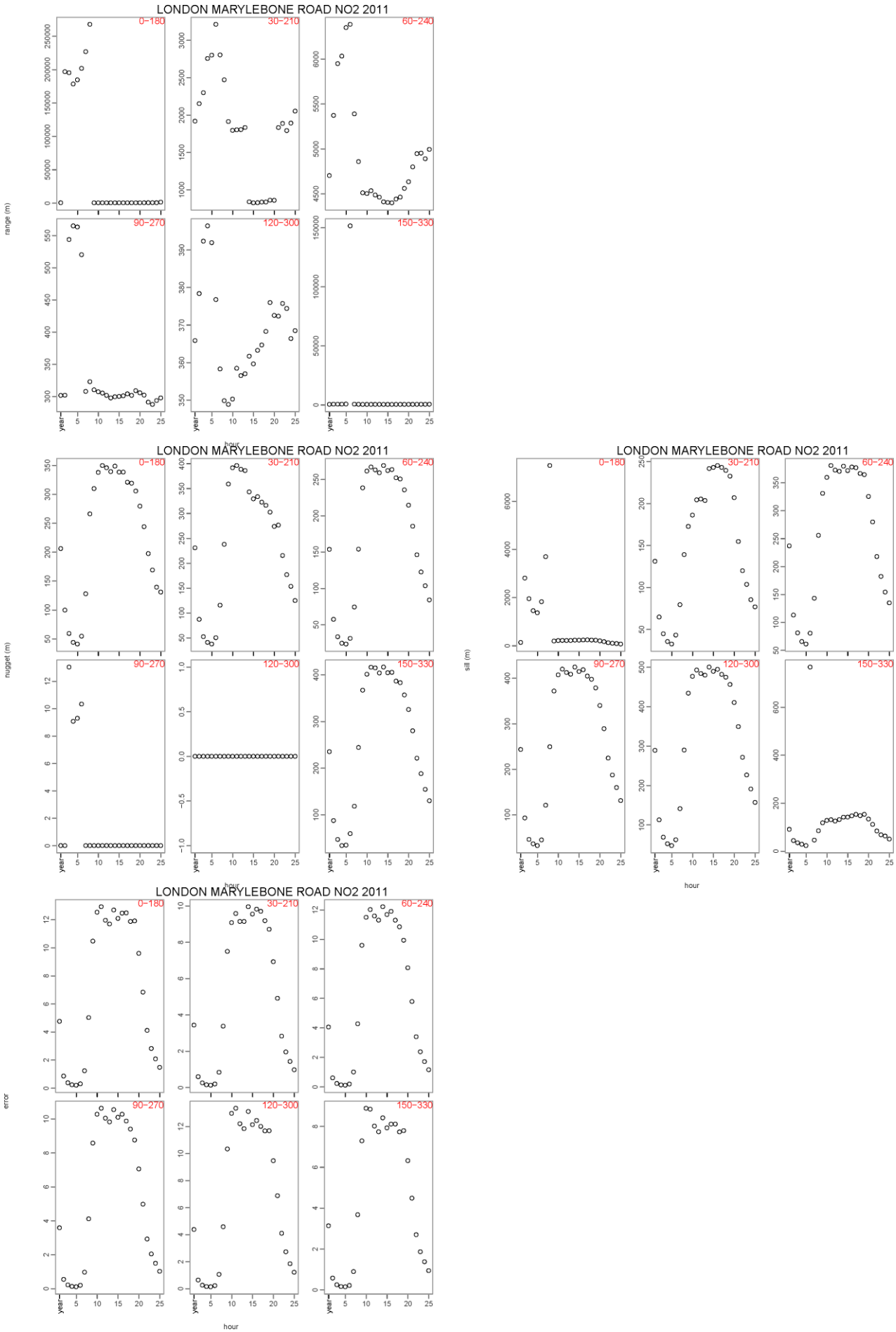


Figure 8. Collective model parameters for the receptor of Marylebone RD, year 2011, species NO_2. Clockwise from top right: range, nugget, error, and sill.

Although the trend of the data for range, nugget and sill is similar for PM_{10} over the two years, there are some quantitative differences, more pronounced for the nugget estimation which is generally lower in

2011 most probably due to the lower variability discussed at the end of Section 3. Also the error is lower for 2011, for the same reason.

By comparing the results of PM_{10} to those for NO_2 (Fig . 8) it emerges a much more pronounced spatial inhomogeneity for NO_2. The directionality plays a stronger role and the variance is largely influences by the hour of the day. Night to early hours (0 to 6) are associated to high values of variance and sudden spikes (see for example the hour 5 for the range in the 150°-330° direction). Errors for NO_2 are also larger than for the PM_{10} case. Although NO_2 and PM_{10} are both primary pollutants emitted by roadside vehicles, the estimation of the fitting parameter using the *pcsv* method differs substantially. Indeed, this has consistently been found when comparing results for the other stations, although a thorough examination of the available results, including O_3, is yet to be performed.

5. CONCLUSIONS AND RECOMMENDATIONS FOR FUTURE WORK

In the framework of the SG1 activity, we have developed a novel technique for assessing the spatial representativeness of ambient air quality monitoring stations and the associated uncertainty. The technique – point-centred semi-variogram (pcsv) – is a modification of existing and well-established geostatistical methods for investigating spatial continuity, such as variogram mapping and modelling. We have applied the technique to an air pollution dispersion modelling product used as proxy of monitored ambient air quality in a busy urban area such as London. Although the results presented so far are preliminary, we showed that the pcsv methodology can help addressing the problem of determining the area of representativeness around a monitoring station with the associated uncertainty.

We believe that, with further refinements and testing, our methodology could be successfully applied to answer scientific and policy related question, such as:

- **Optimal strategy for model evaluation: to what extent comparing a cell-averaged model output to a single point measurements is informative?**
- **Guidelines for optimizing data assimilation in emergency and forecasting models**
- **Spatial uncertainty associated to measurements, which, to date, is still debated.**
- **Using the variogram method, estimate the directional area of representativeness and the maximum distance satisfying the EU data quality objective: for example, the EU directive on data quality objective for NO_2 requires a maximum uncertainty of 25% of the limit value of 100 ppb (any average)). The pcsv model provides readily available variance estimation. It will be therefore straightforward to apply the methodology to answer the EU directive.**

The initial results on the parameters obtained by the variogram model indicate that the spatial continuity around a monitoring site is highly dependent on time, pollutants and direction. Anisotropy (both zonal and geometric) has been found to be predominant in directions at angle with the main street axes.

Known issues to be further investigated:

- **Investigating the causes and solution of trend in the data;**
- **Ensure the robustness of the fitting;**
- **Test the efficacy of the extrapolation method in the vicinity of the receptor;**
- **Remove the artificious directional symmetry introduced by the squared distance by analysis individually sectors of 30° interval;**
- **Dealing with outlying values that mask the goodness of the fit.**

REFERENCES

Bobbia, M., Cori, A., De Fouquet, C., 2008. Représentativité spatiale d'une station de mesure de la pollution atmosphérique. Pollution Atmosphérique N°197.

Carruthers D.J., and et al. 2001. Determination of compliance with UK and EU air quality objectives from high resolution pollutant concentration maps calculated using ADMS-Urban. Int. J. Environment and Pollution, 16, no. 1-6, 460-471.

Isaaks, E.H., Srivastava, R.M., 1989. An introduction to applied geostatistics. Oxford University Press, Oxford (UK)

Janis, M.J., Robeson, S.M., 2004. Determining the spatial representativeness of air-temperature records using variogram-nugget time series. Physical Geography 25, 513-530.

Appendix 1 - Point-centered semi-variogram algorithm**Error! Not a valid link.** As requested during the opening meeting, the code underlying the analysis is produced in R for which the package `gstat` provides ready-to-use variogram tools, such as the `variogram` function, central to this analysis. The code `main_v05.02.r` (a copy of which is attached to this document) performs variogram analysis for the receptors of Table 1 within the modeled domain of London. The modeled grid is treated as grid of receptors around the monitoring point treated as centre of a circle of radius R = 5 km. Thus, the standard variogram cloud is computed (pair-wise distance for all possible combinations); the cloud is then processed to extract only the distances between the central point and all of the other. This reduced cloud is then fitted to an exponential curve whose properties are summarised in a table.
In brief, the various the main code:
1. reads in the monitoring and model data;
2. for each of the receptors of Table 1 selects the modeled concentration data within a distance $\leq R$.
3. calculates the variogram cloud for all pair of points (`CLOUD` option must be on);
4. post-processes the cloud of distance to select only the points relative to the receptor;
5. produces model fit to the variogram using the spherical option,
6. makes the plot and write the results of the fit out to a table.

Auxiliary programs, created ad-hoc for this analysis, perform model fitting (`autofit_fitting_of_variograms.r`) and produces collective plots of the fitting parameters (`plotting_par.r`).

A few remarks

- Due to the large number of grid points within each radius (in excess of 500,000) the irregular model grid was first interpolated to a uniform one (50 m spacing) to facilitate computer time. This is a technical aspect and has no consequence as for the development of the code which works on both regular and irregular points. However, it is important to point out *that the methodology of producing the variogram cloud does not seem feasible for number of points in excess of 100,000 on a single processor machine*. After the interpolation, each receptor is central to a circle of ~61,000 points. These order of magnitudes should have no impact for application to "real world" monitoring networks.
- The only way to use the `variogram` function of the `gstat` package for our purposes is to produce the full cloud first. This produces an output table where all individual point pairs (up to some spatial separation distance) are reported, and identified by columns "left" and "right". Running the `variogram` function with `CLOUD` set to TRUE creats an object of class `variogramCloud`, with the field *np* encoding the numbers of the point pair that contributed to a variogram cloud estimate, as follows. The first point is found by 1 + the integer division of *np* by the `.BigInt` attribute of the returned object, the second point by 1 + the remainder of that division. `as.data.frame.variogramCloud` does the decoding into:

- Left: for variogramCloud data id (row number) of one of the data pair
- Right: for `variogramCloud` data id (row number) of the other data in the pair

Therefore, say v[[m]] the full variogram for the i^{th} receptor, the i^{th}-receptor centered variogram vm[[i]] is found as:

```
vm[[i]]  <-  v[[i]][((v[[i]]$np%/%  attr(v[[i]],".BigInt"))+1)==  1  |  ((v[[i]]$np%%
attr(v[[i]],".BigInt"))+1)== 1, ]
```

This technique is a variation of what is suggested in the R-sig-geo forum by the `gstat` developer (http://r-sig-geo.2731867.n2.nabble.com/Is-it-possible-to-calculate-the-semivariances-of-a-variogram-between-one-specific-point-and-all-othe-td7580909.html#a7580910)

- In its current formulation the code requires that the receptor coordinates are entered as the first elements of the grid matrix, and this explains the "== 1" in the code's line above. However, to deal with a network of *m* receptors indexed as *r(m)* it would be enough to set "== *r(m)*" in a loop over *m* to obtain the variogram centered on each point of the network.
- On the options of the `variogram` function:

There are two important arguments to variogram that, in my opinion, require sensitivity testing: width and cutoff. The argument width seems to means equal-width binning. Providing our own value of cutoff excludes the larger distances. Unfortunately, I have not been able to find in the gstat documentation what happens when both cutoff and/or width are provided in addition to boundaries. This latter option is passed as a named argument to the function variogram during the fitting process. By looking at the source code for autofitVariogram, boundaries is defined as the distance intervals which define the bins, and is set as 2%, 4%, 6%, …, 100% of about 1/3 the diagonal of the box spanning the sample locations (defined using the spDists function from the sp package). Options to test would be:

- Create bins of equal elements (populated evenly)
- Create bins of equal width (possibly done by the width option, but what about boundaries?)

Code for running these tests should be easy to produce.

- The strategy of producing the cloud for all combinations of point pairs to select only those of interest is not computationally sound. It has been dictated by the availability of the fitting function autofitVariogram of the automap package, which requires an object of class gstat as input. Indeed, the computational time mainly depends on the production of the cloud. An independent script has been developed that computes the receptor-centered semivariogram and which makes no use of predefined functions. For testing purposes this script have been firstly developed in MatLab and then translated into R. The results have been compared with those of the variogram functions discussed above to make sure there are no discrepancies. The MatLab code is reported hereafter:

```
dat = importdata('D:/work/spat_repres/code/input/recs_for_matlab.csv',',')
X = dat.data(:,2);
Y = dat.data(:,3);
Z = dat.data(:,1);
N = length(Z);
%Calculat PointCenteredSemivariogram(X,Y,Z,N)
for i=1:N
    for j=1:N
        Distance(i,j)=sqrt((X(i)-X(j))^2+(Y(i)-Y(j))^2);
    end
    %
    % Half-squared difference summation of variable
    %
    for j=1:N
        HalfSquareDiff(i,j)=0.5*(Z(i)-Z(j))^2;
    End
end
    %
    % Distance-Half distance matrix
    DisHal=[Distance(i,1:N);HalfSquareDiff(i,1:N)];
```

This script is simple and faster to run with respect to the one discussed above. However, to exploit the full functionality of the variogram fitting function available in R a gstat object is required, which is directly produced by the variogram function.

In summary, we have the options of *i)* continuing with processing the full cloud (computationally intensive) and using the available fitting functionality or *ii)* using the simple script above (fast to run) and develop our own fitting functions (time consuming in testing and debugging).
After discussing these options with M. Gerboles it was decided to stick with the variogram functionality and use the built in autofitvariogram function.

The autofitVariogram function automatically fits a variogram to the data on which it is applied. The automatic fitting is done through fit.variogram. In fit.variogram the user had to supply an initial estimate for the sill, range etc. autofitVariogram provides this estimate based on the data and then calls fit.variogram. A few simple choices are made when estimating the initial guess for fit.variogram. The initial sill is estimated as the mean of the max and the median of the semi-variance. The initial range is defined as 0.10 times the diagonal of the bounding box of the data. The initial nugget is defined as the min of the semi-variance. There are five different types of models that are often used:

Sph : A shperical model; Exp: An exponential model; Gau: A gaussian model; Mat: A model of the Matern family; Ste: Matern, M. Stein's parameterization

We have applied out technique by applying the Sph option only.

European Commission
EUR 26539 EN – Joint Research Centre – Institute for Environment and Sustainability

Title: **Spatial representativity - Report of 2013 WG2/SG1 activity**

Author(s): Efisio Solazzo, Michel Gerboles, Oliver Kracht, Jenny Stocker*, David Carruthers* and Stefano Galmarini
(*) Cambridge Environmental Research Consultants (CERC), Cambridge (UK)

Luxembourg: Publications Office of the European Union

2014 – 21 pp. – 21.0 x 29.7 cm

EUR – Scientific and Technical Research series – ISSN 1018-5593 (print), ISSN 1831-9424 (online)

ISBN 978-92-79-35592-9 (print)
ISBN 978-92-79-35591-2 (pdf)

doi: 10.2788/14035

Abstract

The attached document addresses the progress of the FAIRMODE/SG1 activity since it starts in June 2013 up to the month of November 2013, for a total of about 100 working days.
During this period the analysis has focused on the development of a novel methodology – *point-centred semi variogram* (pcsv) – for help assessing the spatial representativeness of the air quality receptors in Europe. The technique has been successfully developed and tested on a proxy of measured ambient air quality concentration data, namely results from an atmospheric modelling system. Action taken:

1. Established collaboration with the CERC group of the UK, developer and distributor of the Atmospheric Dispersion Modelling System (ADMS), for sharing of model results of pollutant dispersion in London. These data have been used as proxy in our analysis;
2. Established collaboration with the ARIA technologies SA of France for sharing of the AIRCITY high resolution modelling results of pollutant dispersion in Paris. The ARIA group has manifested the willingness to cooperate on the SG1 activity, although data for Paris have not yet been exploited.
3. Development, testing, and coding of the pcsv technique;
4. Application of the pcsv methods to the ADMS data using the positioning of European AIRBASE air quality receptors as centre of the analysis;
5. Preparation of the attached report.
6.

We note that about 90% of the total time has been spent on item 3 and 4 of the action list. The code we have produced is computing-intensive and has been run at the limit of the available computing resources (single PC unit). The code requires about ten days to complete a full run (three species, two years, fourteen central points). The results obtained are promising, the technique is robust and applicable to several pollutants, though it should be considered that generalisation is difficult, as any other cases needs to be treated individually. The necessity of developing a generalised method will require more time as other cases will need to be considered and the analogies among them taken in to account.

LB-NA-26539-EN-N

JRC Mission

As the Commission's in-house science service, the Joint Research Centre's mission is to provide EU policies with independent, evidence-based scientific and technical support throughout the whole policy cycle.

Working in close cooperation with policy Directorates-General, the JRC addresses key societal challenges while stimulating innovation through developing new methods, tools and standards, and sharing its know-how with the Member States, the scientific community and international partners.

Serving society
Stimulating innovation
Supporting legislation

Publications Office

ISBN 978-92-79-35591-2